Adversarial autoencoders for novelty detection

Valentin Leveau, Alexis Joly

ADVERSARIAL AUTOENCODERS FOR NOVELTY DETECTION

Valentin Leveau
INRIA Zenith
vleveau@inria.fr

Alexis Joly
INRIA Zenith
ajoly@inria.fr

ABSTRACT

In this paper, we address the problem of novelty detection, *i.e* recognizing at test time if a data item comes from the training data distribution or not. We focus on Adversarial autoencoders (AAE) that have the advantage to explicitly control the distribution of the known data in the feature space. We show that when they are trained in a (semi-)supervised way, they provide consistent novelty detection improvements compared to a classical autoencoder. We further improve their performance by introducing an explicit rejection class in the prior distribution coupled with random input images to the autoencoder.

1 INTRODUCTION

Supervised deep learning architectures have been successfully used in a wide range of object classification tasks showing impressive results while discriminating between a lot different visual concepts. However, as powerful as such models are, one issue is that they are optimized to predict from a restricted set of categories. So that, if a new data item belonging to an unknown category comes at the prediction phase, the model systematically predicts a known label with a possibly high confidence. In a practical application, we rather would like to detect that the item does not belong to any of the known classes so as to inform the user about this mismatch. This fundamental *novelty detection* problem has received a lot of attention in the machine learning community (a comprehensive review can for instance be found in (Pimentel et al., 2014)). Novelty detection methods can be roughly classified into probabilistic approaches (parametric or nonparametric), distance-based approaches (e.g. nearest neighbors-based or clustering-based), reconstruction-oriented approaches (e.g. neural network-based or subspace-based) and domain-based approaches (e.g. one-class support vector machines). In this paper, we chose to investigate how unsupervised and semi-supervised deep learning methods could help solving the novelty detection problem. Basically, such methods (Hinton et al., 2006; Bengio & LeCun, 2007; Vincent et al., 2010; Rifai et al., 2011; Goodfellow et al., 2014; Makhzani et al., 2015) aim at disentangling and capturing the explanatory factor of variation of the data. This can be seen as modeling the data generating distribution/process. By doing so, we expect the system to learn the manifold on which the training data lies and to generalize on it. By extension, we expect that new data items belonging to unknown classes won't be well captured and that the generative model will fail to reconstruct them accurately. In this paper, we focus in particular on Adversarial Autoencoders (AAE) (Makhzani et al., 2015) that have the advantage to explicitly control the distribution of the known data in the feature space, so that it is possible to quantify the likelihood that an image belongs to the manifold of the known training data. We explore the use of both unsupervised and supervised prior distributions and we introduce a new variant that explicitly models a rejection class in the latent space. Experiments on MNIST dataset show that this variant provides better novelty detection performance than classical autoencoders and adversarial autoencoders.

2 PROPOSED NOVELTY DETECTION METHODS

Baseline: reconstruction-based novelty detection through autoencoders: Using the reconstruction error of a generative model is a well known novelty detection method (Pimentel et al., 2014; Thompson et al., 2002). The higher the reconstruction error of an item is, the farther from the manifold of the known training data it is expected to be. As a baseline novelty detection method, we thus suggest to use the reconstruction error of a (deep) autoencoder. The autoencoder we used in our

experiment for evaluating this baseline is the one described in (Makhzani et al., 2015) with 3 fully connected layers for both the encoder and decoder. The ReLU activation function is used for all layers except for the last layer of the encoder that is linear and the last layer of the decoder that uses a sigmoid activation function. The reconstruction error is the Euclidean distance between a sample of the input space and its reconstruction and is directly used as the novelty detection criterion:

$$\rho_1(\mathbf{x}) = ||\mathbf{x} - g(f(\mathbf{x}))||_2^2 \tag{1}$$

where $f(\mathbf{x})$ and $g(\mathbf{x})$ are respectively the encoding and decoding function of the autoencoder.

Adversarial Autoencoders for Novelty Detection: An adversarial autoencoder (AAE) is a probabilistic autoencoder that uses the recently proposed generative adversarial networks (GAN) to perform variational inference by matching the aggregated posterior of the hidden code vector of the autoencoder with an arbitrary prior distribution (Makhzani et al., 2015). The decoder of the adversarial autoencoder learns a deep generative model that maps the imposed prior to the data distribution. In this paper, we explore in what way AAEs might be useful for the novelty detection problem. Therefore, we define a new novelty detection criterion based on the likelihood of a candidate sample according to the imposed prior:

$$\rho_2(\mathbf{x}) = 1 - p(\mathbf{f}(\mathbf{x})) \tag{2}$$

where $p(\mathbf{z})$ is the imposed prior distribution, *i.e.* the higher $p(\mathbf{f}(\mathbf{x}))$ and the more likely $\mathbf{x}$ belongs to the training data distribution. In our experiments, we focus on two prior distributions:

- *Normal distribution*: $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$. This is the default distribution used in Makhzani et al. (2015) to ensure that generating from any part of the prior space results in meaningful samples.
- *Gaussian mixture*: $p(\mathbf{z}) = \sum_i p(\mathbf{z}|C_i)$ with $p(\mathbf{z}|C_i) = \mathcal{N}(\mu_\mathbf{i}, \mathbf{\Sigma_i})$. This is the prior distribution suggested in to handle supervision or semi-supervision. To ensure the mapping between the labels of the training data items and the classes of the Gaussian mixture, it is required to pass as input of the adversarial discriminator a one-hot vector coding the label of $\mathbf{z}$ in addition to $\mathbf{z}$ itself. Complementary to this *Supervised Gaussian mixture* prior, in our experiments, we also evaluated the case of an *Unsupervised Gaussian mixture* by removing the label's one-hot vector from the input of the adversarial discriminator. This allows us to evaluate separately the benefit of the Gaussian mixture (over the normal distribution), and the benefit of the supervision.

Adversarial Autoencoder with an explicit rejection class: It is important to notice that the likelihood $p(\mathbf{f}(\mathbf{x}))$ of an item $\mathbf{x}$ does not model the real probability that it belongs to the manifold of the known data. Considering this likelihood as a probability is in fact a case of prosecutors fallacy since $p(\mathbf{f}(\mathbf{x}))$ should be rather noticed $p(\mathbf{f}(\mathbf{x})|y(\mathbf{x}) = 1)$ where $y(\mathbf{x})$ is a binary function indicating whether $\mathbf{x}$ belongs to one of the known classes or not. Then, what we would like to estimate is rather

$$\rho_3(\mathbf{x}) = p(y(\mathbf{x}) = 0|\mathbf{f}(\mathbf{x})) = \frac{p(\mathbf{f}(\mathbf{x})|y(\mathbf{x}) = 0).p(y = 0)}{p(\mathbf{f}(\mathbf{x})|y(\mathbf{x}) = 1).p(y = 1) + p(\mathbf{f}(\mathbf{x})|y(\mathbf{x}) = 0).p(y = 0)} \tag{3}$$

But since we don't know anything about the conditional likelihood of the unknown data $p(\mathbf{f}(\mathbf{x})|y(\mathbf{x}) = 0)$ (and about the novelty rate $p(y = 0)$), we can not estimate that probability. To try overcoming this issue, we propose to explicitly add a novelty class to the prior distribution of the AAE. More precisely, we model the unknown data by a normal distribution at the center of the latent space (*i.e.* $p(\mathbf{f}(\mathbf{x})|y(\mathbf{x} = 0) = \mathcal{N}(0, I)$) and we still model the known data by a mixture of non-centered Gaussian functions $p(\mathbf{f}(\mathbf{x})|y(\mathbf{x} = 1) = \sum_i p(\mathbf{z}|C_i)$. Then, we enforce the autoencoder to map the unknown data space onto $p(\mathbf{f}(\mathbf{x})|y(\mathbf{x} = 0) = \mathcal{N}(0, I))$ by adding to the training set some random images and by passing the known/unknown label $y(\mathbf{x}) \in 0, 1$ as input to the discriminator as any other class label.

3 EXPERIMENTS

Protocol and settings: We evaluated the novelty detection methods described above on MNIST handwritten digit dataset (LeCun et al., 1998) (10 classes, each with approximately 600 images for training and validation and 100 images for testing). To fit with our novelty detection scenario, we removed from the training set 3 of the 10 classes (the '2', '5' and '7' digits). We then computed

our novelty detection criteria (ρ_1, ρ_2 and ρ_3) on the entire test set and we measured the performance through a Mean Average Precision (mAP). A mAP that is equal to 1 means that all the test images of digits '2', '5' and '7' (unknown classes) have a higher ρ value than the images of the known digits. All autoencoders have been trained through back-propagation using the Nesterov momemtum solver (Sutskever et al., 2013) using a learning rate and a momentum parameter respectively set to 0.1 and 0.9. We iterated over 2000 epochs (without validation) using mini-batches of 128 images. For the AAE, each epoch includes 3 steps: (i) reconstruction optimization phase, (ii) discriminator optimization phase and (iii) generator optimization phase.

Results: Table 1 and Figures 1 and 2 provide the results of our experiments when using a 2-dimensional latent space for the autoencoders. This is of course not an optimal feature dimension in terms of performance but this allows visualizing how the known and unknown test samples are distributed in the latent space. The results first show that fully unsupervised Adversarial Autoencoders do not perform better than baseline autoencoder (using the reconstruction error criterion ρ_1). Looking at Fig1 (b) and Fig1(c), the main reason is that the unknown samples are mapped according to the same prior distribution than the known samples. As stated in section 2, it is actually not because the known data items are enforced to lie in the dense regions of the prior, but that any data item in such regions is belongs to the manifold of known data.
The second major conclusion is that the AAE using a *supervised* GMM prior clearly outperforms the baseline autoencoder (contrary to the AAE using the *unsupervised* GMM prior). As shown in Fig1 (d), the addition of the supervision enforces the encoder to map the unknown data items away from the known classes, and, by default, at the center of the feature space. Actually, the center of the latent space seems to act as an attractor of the default open space. This might be related to the fact that, whatever the used prior distribution, randomly generated images are distributed according to a normal distribution at the center of the feature space (because of the central limit theorem).
The third conclusion of this preliminary study is that the likelihood-based and posterior-based novelty detection criteria are less effective than the reconstruction error criteria. But this has to be mitigated for several reasons. First, this might be specific to the MNIST dataset for which the original image space is already very well shaped so that the L_2-distance between an image and its reconstruction is semantically meaningful. But this might not be the case for more complex data that would require to capture more invariance and spatial structures (*e.g.* using ConvNets). We can expect that the likelihood-based and posterior-based criteria would be less sensitive to such higher complexity than the reconstruction error. Another advantage is to enable a normalized and well interpretable novelty score, in particular the posterior-based criterion that is a real probability.

Representation Learning Methods		Novelty detection criterion		Figures (Appendix)
		reconstruction error (ρ_1)	likelihood (ρ_2) or posterior (ρ_3)	
autoencoder		0.71	-	1(a) and 2(a)
Adversarial autoencoder	*Normal distribution*	0.68	0.35 (ρ_2)	Fig 1 (b)
	Unsupervised GMM	0.64	0.41 (ρ_2)	Fig 1 (c)
	Supervised GMM	0.82	0.82 (ρ_2)	Fig 1 (d)
Adversarial autoencoder + rejection	*Supervised GMM*	**0.89**	**0.83** (ρ_3)	Fig 1 (e)

Table 1: mAP of the different novelty detection methods

4 CONCLUSION

In this preliminary study, we investigated the use of Adversarial autoencoders for the hard problem of novelty detection. We did show that imposing a supervised prior distribution can help mapping the unknown items away from the known classes but that it is still theoretically not possible to control their distribution in the feature space. Overall, we believe this remains an open question that requires to first understand whether novelty should be conceptualized as unusual recombination of elements of prior knowledge or not.

REFERENCES

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

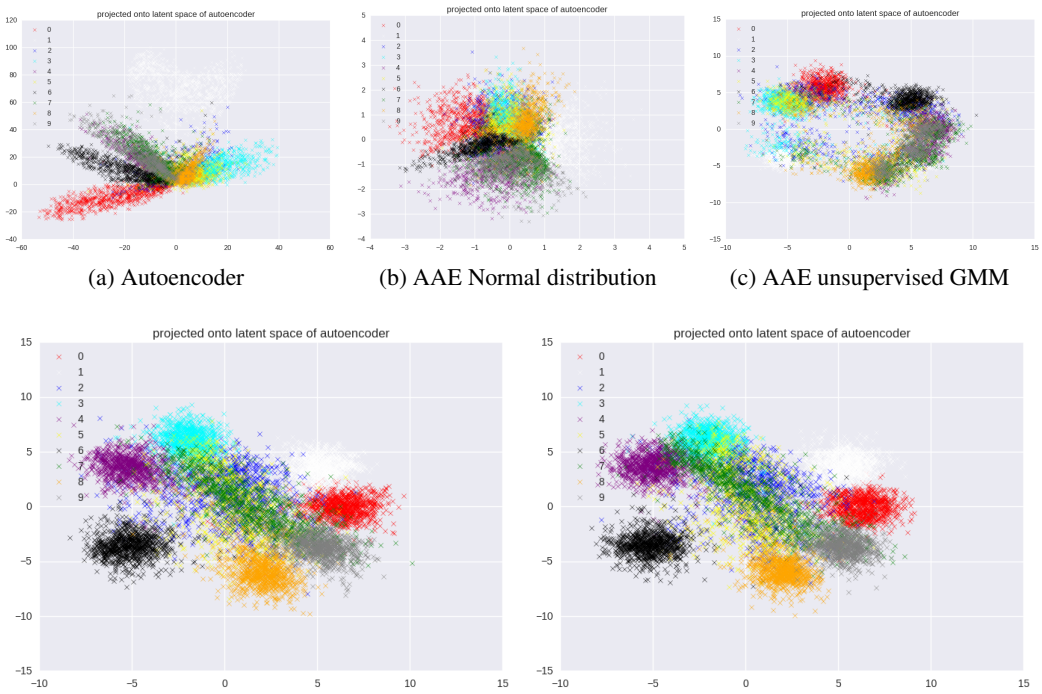
Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 833–840, 2011.

Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. *ICML (3)*, 28:1139–1147, 2013.

Benjamin B Thompson, Robert J Marks, Jai J Choi, Mohamed A El-Sharkawi, Ming-Yuh Huang, and Carl Bunje. Implicit learning in autoencoder novelty assessment. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, volume 3, pp. 2878–2883. IEEE, 2002.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

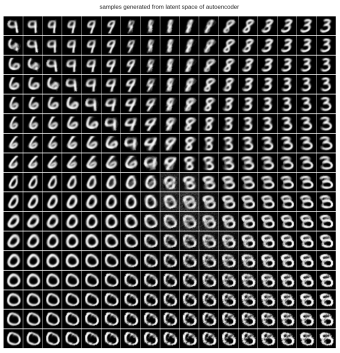
APPENDIX



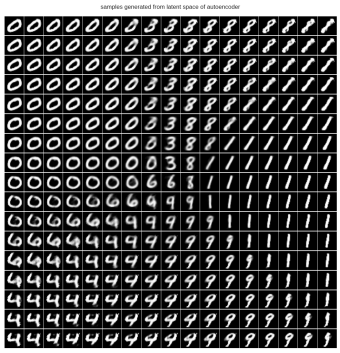
(a) Autoencoder (b) AAE Normal distribution (c) AAE unsupervised GMM

(d) AAE supervised GMM (e) AAE supervised GMM+reject

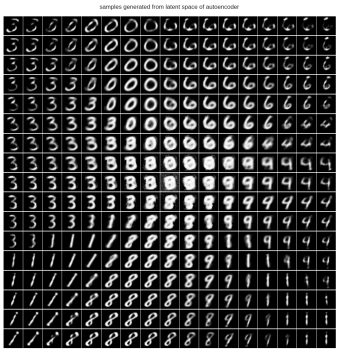
Figure 1: Visualization of the test samples in the latent space (unknown digits are '2' (dark blue), '5' (yellow) & '7' (green))



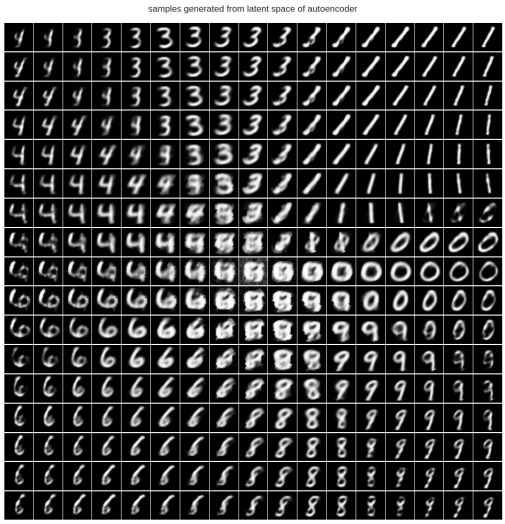
(a) Autoencoder



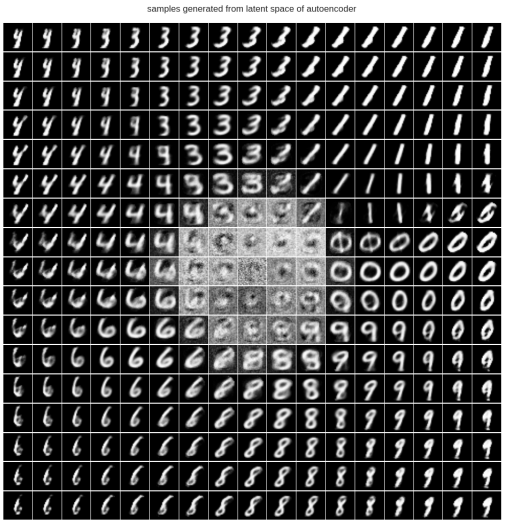
(b) AAE Normal distribution



(c) AAE unsupervised GMM



(d) AAE supervised GMM



(e) AAE supervised GMM+reject

Figure 2: Visualization of the reconstructed latent spaces plotted on Figure 1. Images were obtained by uniformly sampling vectors in the latent space and by feeding them to the decoder function. When no supervision is used, the whole test set is learned to be reconstructed as images coming from classes of the training set. We can see in Fig (2e) that learning to reconstruct the noisy images coming from the rejection class while using the supervision to shape the latent space allows us to push the images of the novel classes toward the regions that are reconstructed as noisy images.